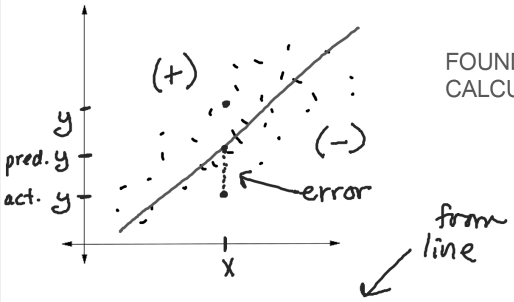
Chapter 8: **Line of Best Fit**



FOUND ON CALCULATOR

Errors = *residuals* = actual Y -- predicted Y

line of best fit ~ average (mean) of the plot

sum of + errors = sum of -- errors

Linear Regression Line:

- Describes how… a response variable (Y) changes as an explanatory variable (X) changes (#)
- Used to … predict the value of Y for a given value of X

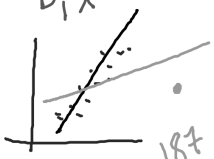
$Y = mX + b$

10 mg

Most accurate Regression line:

- Called: Least Squares Regression Line (LSR line or LSRL)
- Definition: minimizes… the $(\text{errors})^2$ of Y
- Form: $\hat{y} = b_0 + b_1x$ (hat = sample; int.; slope)

$b_1 = r\left(\frac{S_y}{S_x}\right)$ $b_0 = \bar{y} - b_1\bar{x}$



p.187

- always … passes thru $(\bar{X}, \bar{Y})$
- not resistant (to outliers) = outliers affect the line a lot!
- on calculator: STAT -> CALC -> 8:LinReg(a + bx) XLIST, YLIST, Y1

***note: you get Y1 by going VARS--> YVARS-->FUNCTION -->Y1*

Calculator Example (using EXA1 and EXA2)

$(\widehat{2^{nd}\text{exam}}) = 24.86 + 0.6702(1^{st}\text{exam})$

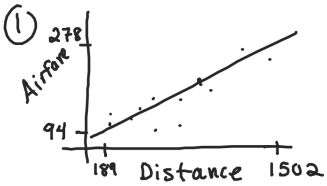
$\hat{y} = 24.86 + 0.6702(x)$

$x = 1^{st}\text{exam}$

$y = 2^{nd}\text{exam}$

Complete worksheet 8A

Worksheet 8A Answers:

① 

- mod. strong
- linear
- positive
- poss. outliers @ (612,94) and (737,98)

②

	mean	Std. Dev.	corr
x	712.667	402.69	0.795
y	166.92	59.45	

③ $b_1 = r\left(\frac{S_y}{S_x}\right)$ $b_0 = \bar{y} - b_1\bar{x}$ (no LinReg)

③ $b_1 = r\left(\frac{s_y}{s_x}\right) = 0.795\left(\frac{59.45}{402.69}\right) = 0.1174$

$b_0 = \bar{y} - b_1\bar{x} = 166.92 - (0.1174)(712.667)$

$= 83.25$

$\hat{y} = 83.25 + 0.1174x$

Airfare = 83.25 + 0.1174 (distance)

④ Airf = 83.267 + 0.1174 (dist) $r = 0.795$

⑤ $\hat{y} = 83.267 + 0.1174(370) = \$126.70$

Error = actual − predicted

= 138 − 126.70

= \$11.30

*predictions on calculator

$\hat{y} = 83.267 + 0.1174(1502) = \$259.56$

Error = 258 − 259.56 = −\$1.56

⑥ $\hat{y} = 83.25 + 0.1174(2842)$

= \$416.85

⑦ Error = 198 − 416.85 = −\$218.85

⑧

D	900	901	902	903
A	188.90	189.02	189.14	189.26

⑨ \$0.12 = slope

⑪ \$12.00

\$11.74

Interpreting the slope:

Sentence: For every increase of 1 x-variable (units) the y-variable increases/decreases by slope (units) on average

(+ / −)

Example: Airfare data. Slope = 0.1174 dist vs. airf.

slope = change in y = \$0.1174 / change in x = 1 mile

For every 1 mile flown, the airfare increases by \$0.12 on average.

Another example:

The LSRL for MPG vs. Horsepower for different models of cars is:

$\hat{MPG} = 46.87 - 0.084(HORSEPOWER)$

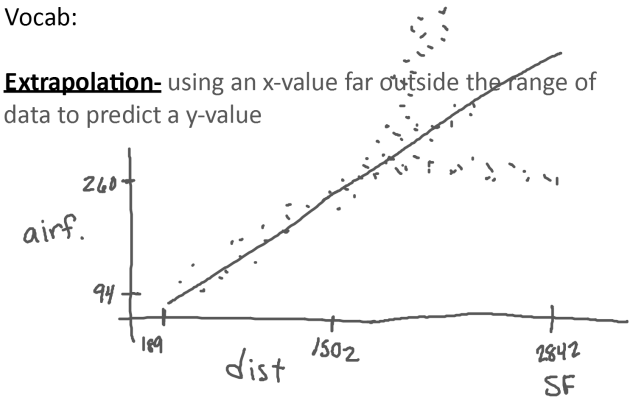
Interpret the slope:

$\frac{-0.084 \text{ mpg}}{1 \text{ hp}} = \frac{y}{x}$

For every increase of 1 horsepower in a car, the mpg decreases by 0.084 mpg on average.

Vocab:

Extrapolation- using an x-value far outside the range of data to predict a y-value



inaccurate / unsure of reliability

Coefficient of determination:

symbol & calculation: r^2 ** listed as a percent

variability $r^2 = 0.89235 \rightarrow 89.235\%$

sentence interpretation:

explained by

r^2 % of the **change** in the y-variable is due to the **change** in the x-variable. (*or* due to the LSR line)

Example: For airfare data, r = 0.795, so $r^2 = 0.632$

63.2%

63.2% of the change in the airfare is due to the change in the distance flown.

RESIDUALS: Example 1 **Ungroup RESID (lists LIST1 and LIST2)**

1. Create a scatterplot of LIST1 (x) vs. LIST2 (y). Describe the plot.
2. Find the LSR line, r, and r2. Add the LSR line to your plot.

3. Look at the plot. Do you think that the line does a good job of describing the trend of the data?
4. Does it hit every point though?

5. There are obvious errors/residuals. What type of residuals will the points that fall ABOVE the line have? positive or negative?
6. What type of residuals will the points that fall below the line have?

7. Will all the residuals be the same number? or will they all be varied?

8. What can you say about the number of positive vs. negative residuals? Will there be about the same amount? Or more of one type?

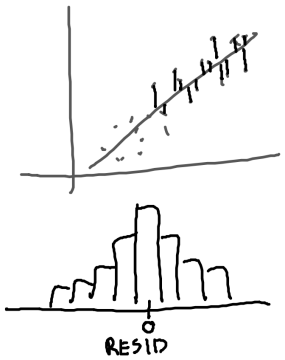
Residuals (errors):

= actual y -- predicted y

$= y_i - \hat{y}_i$

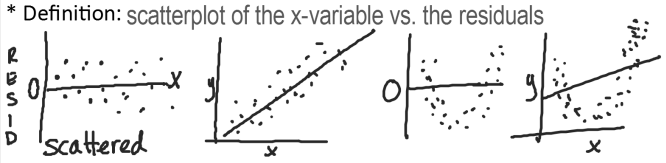
sum of residuals = 0

mean of residuals = 0



Residual Plot

* Definition: scatterplot of the x-variable vs. the residuals



* Helps... assess the fit of the LSR line
* No pattern = scattered = our line is a good model for the data
* Pattern = another model (quadratic, exponential, etc.) would be a better fit for the data ~~line is a bad fit~~

Examples:

On Calculator:

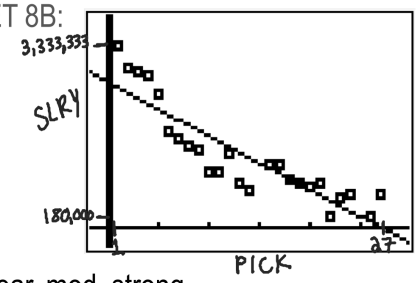
Scatterplot
Xlist: *keep your same X-variable*
Ylist: $_L$RESID

NOTE: Must run LSR line first!! (LinReg)

Complete worksheet 8B

WORKSHEET 8B:

1)



Negative, linear, mod. strong
(or slight curve)

2) $\widehat{SALARY}$ = 2,657,443.08 - 98,957.22(PICK)

r = -0.8869 r^2 = 0.7866

3) r^2 = 78.66%

Y1(12)

4) $\hat{y}$ = 2657443.08 - 98957.22(12) = $1,469,956.49

residual = $1,370,000 - $1,469,956.49 = -$99,956.49

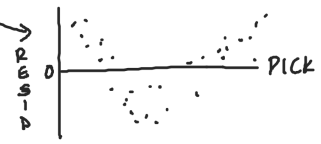
overestimate

5) $\hat{y}$ = 2657443.08 - 98957.22(15) = $1,173,084.84

$\hat{y}$ = 2657443.08 - 98957.22(25) = $183,512.68

6) slope = $98,957.22

7) the linear model does not appear to be the best model for the data



8) sum = approx. 0

9) +resid = underestimate (actual > predicted)

10) - resid = overestimate (actual < predicted)

11) no outliers